Planets - Preservation and Long-term Access through NETworked Services

Outreach and Training events 2009-10: 'Digital Preservation – the Planets Way': summaries for technical / developer staff

Raymond van Diessen (IBM), Laura Molloy and Andrew McHugh (both HATII, University of Glasgow)

Contents

1. Introduction to Digital Preservation: Why Preserve? How to Preserve?

Watch this presentation at:
http://www.planets-project.eu/training-materials/1-king-planets_keynote/

The world is quickly becoming digital. In 2007, it was estimated that the total amount of digital information was around 281 exabytes[1]. At the end of 2009, the estimated amount is around 700 exabytes – this represents an increase of 60% over only two years!

For the first time, the amount of digital information produced exceeds the storage space available. Naturally, not all of this information has to be maintained for the long term, but unarguably more and more born-digital information will be created and attention needs to be paid to preservation for the long-term. Unlike information stored on physical media such as stone, paper or parchment, information in digital media cannot be read directly by the human eye. Access requires the information, encoded as bits and bytes according to the rules of a particular format (the context), to be processed by software and physically displayed.

As the growth rate of digital information continues to increase exponentially, technology innovation also continues to advance. There are not many people that work with the same computer for more than four to five years. Every year, new file formats or new versions of file formats are being introduced. Will we be able to still listen to our MP3 songs or look at our JPEG holiday photos twenty years from now? It is hard to keep up with both software and hardware changes and if not managed proactively, we can be confident that, twenty years from now, our digital information may not be usable.

This feeling is exemplified in Jeff Rothenberg's famous 1997 quote: "Digital documents last forever – or five years, whichever comes first."[2]

[1] Gantz et al, 2008. *The Diverse and Exploding Digital Universe*, IDC White Paper, International Data Corporation. Available at http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

The long-term preservation challenge can be divided in two aspects:

1. **Bit-stream preservation**: How to keep the bit-stream readable when the storage devices that initially stored them, or the hardware to which the storage is attached, become culturally or technologically obsolete;
2. **Logical preservation**: How to deal with the obsolescence of the software stack (operating system and application software) and the absence of the required context to interpret the information.

Today's digital media have a limited life-span (**media obsolescence**); for example, the 5¼-inch floppy disk (with a 10-year life-span. Other technologies like microfilm (500 years) have a better track record and acid-free paper can have a life-span of 200 years. The bit-streams can only be preserved for the long-term when they are copied to current storage devices and hardware platforms on a regular basis.

A further challenge is to address **format obsolescence** due to changing software and hardware over time and the loss of the needed context to interpret the information. An analogue example of this is Egyptian hieroglyphics. Information is visibly recorded but cannot be read without knowledge of how to interpret the symbols. In order to recognize software and hardware dependencies and to maintain a usable interpretation of the information, the Open Archival Information System Reference Model (OAIS) defined the concept of representation information. **Representation information** refers to all information required to access and understand the information stored within a digital object.

Logical preservation with the aid of representation information defines two primary global approaches:

1. **Emulation**: Uses representation information to re-create the original environment necessary to access the preserved bit-stream, and so express the necessary information to enable any data content to be comprehended;
2. **Migration**: Uses representation information to assist characterisation and validation of files, and to identify endangered file formats and convert them to latest accessible (open, standardised) formats.

The Planets project aims to help users to identify the file format of digital objects and extract specific digital object characteristics (by using the Planets XCL tool and Planets Core Registry). Planets also helps clients address logical preservation in multiple ways, including the introduction of preservation plans, to find out and make decisions on what to preserve and how to do it (cf. Planets preservation tool Plato), and methods to evaluate and execute concrete preservation actions on digital objects (cf. Planets Testbed and preservation action services).

For further information, please see reading list items A1, A2, B1.

[2] Expanded in Rothenberg (1998), *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Council on Library and Information Resources. Available at http://www.clir.org/PUBS/reports/rothenberg/contents.html

2. The Preservation Action Cycle

Watch the presentation at:
http://www.planets-project.eu/training-materials/2-billenness-planets_risk_management/

Many organisations are currently unaware of the risks associated with the long-term preservation of their digital information. Focus is currently often placed on management and accessibility of the digital information for current usage. However, a large part of the born-digital information created today will also need to be usable in the medium- to long-term future. A proven approach for this challenge is to apply standard risk management techniques to long-term digital preservation.

There are a number of risk management process standards available such as the British Standard 31100, 'Risk Management. Code of Practice.'[3] In order to evaluate risk in a consistent manner, a shared context related to long-term digital preservation has to be established. ISO 14721:2003 '**Reference Model for an Open Archival Information System (OAIS)**'[4] successfully provides such a common context. It describes the high-level functional components that should be present in an electronic repository focused on long-term preservation, as well as information components required to support the preservation process.

Each organisation should base its risk management practices on the following well-defined risk management principles:

- Risk management should be part of decision making;
- Risk management should be tailored to fit institutional requirements.

It should be acknowledged that the risk management support implicit in Planets tools and approaches is mainly focused at object and object property level issues, and not wider issues of, for example, organisational, financial and infrastructural sustainability.

The **risk management process** contains five phases and is iterative:

1. **Identify Risk**: during this phase, the risk related to obsolescence for each item in the digital collection is evaluated. The Planets preservation planning tool, Plato, provides a framework to identify and manage the characteristics of digital objects being preserved. These characteristics in turn can be largely automatically extracted by Planets characterisation services for each individual digital object. The Planets Core Registry contains information about characteristics at file format level and the preservation action tools that can be used to preserve them...

2. **Assess Risk**: during this phase the severity, likelihood and immediacy of the risk is assessed. The same tool as in the first phase is used to compare the actual characteristics of each digital object with the profiles managed in Plato.

---

[3] BS 31100:2008. *Risk Management. Code of Practice.* British Standards Institution. Available at http://shop.bsigroup.com/ProductDetail/?pid=000000000030191339

4 ISO 14721:2003. *Space data and information transfer systems -- Open archival information system -- Reference model.* International Organization for Standardization. Available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

3. **Plan to Mitigate Risk**: there are different approaches to preserving digital content and many tools that can be used. During this phase multiple alternatives to address the risk are evaluated. Planets Testbed and Corpora provide the ideal test environment to evaluate various preservation actions in response to a preservation risk. The result of different potential migration and emulation actions can be evaluated and compared to the objectives as defined in the Planets Plato tool. The central installation of the Planets Testbed provides a generally accessible corpora of test digital objects and the associated results of different preservation actions, allowing organisations to make a sound decision on which preservation action to select.

4. **Risk Plan Implementation**: at the heart of risk mitigation is the Planets Interoperability Framework. This provides the glue for the different components to interact with each other, e.g. Plato, characterisation tools, emulators and migration tools, as well as the Planets Core Registry (PCR). The Planets Interoperability Framework is web-based and adopts open standards like J2EE and XML. Individual preservation activities can be linked in configurable business processes, tailored to the needs of the specific institution or collection. The Planets Comparator provides a novel approach to automatically performing a quality assessment of a migration action, by comparing different key characteristics of the original file and the migrated file in a format- and technology-neutral manner.

5. **Risk Plan Review and Update**: the process is a cycle. Based on the execution of the previous four phases, institutional policies and guidelines may be adapted to manage potential risks more effectively or address newly identified risks triggered by an operating system going out of use or a format becoming unsupportable. The PCR is continually updated with information about new formats and tools.

In summary, Planets provides an extensible and integrated environment to manage preservation risks and act on them effectively.

For further information, please see reading list items B2, C1.

3. How to Understand Files

Watch this presentation at:
http://www.planets-project.eu/training-materials/4-schnasse-understanding_files/

Digital content is different from analogue content in that the bits and bytes that encode the content are useless without the right software programs and format standards to interpret them, and hardware to render the content. Take a book as an example of analogue content: to interpret the content, the user opens the book and starts reading. Naturally, the specific interpretation is always dependent on the context, background, capabilities or knowledge base of the reader. This holds true for both analogue and digital content. No other additional support is needed in the case of the analogue book example.

However, to preserve digital content we have to address a minimum of three fundamental aspects (as there may be wider cultural, financial, political, spatial or sociological factors that influence value or perceptions of digital content. Each can have implications for identifying and mitigating preservation risk.) The three fundamental aspects, however, are:

- the content file;
- the software to interpret the content files (processing);
- the hardware that displays the content (representation).

It is important to the viewer that properties they value are preserved and the file is as close to the original as possible. Key to the preservation effort is the ability to characterise / quantify all three aspects to evaluate the success of preserving the digital content. The easiest way to do this would be to compare the resulting presentations of both the original digital content and the preserved version by viewing them side-by-side to see how far they are similar or different. However, this is not a sustainable solution because of the large volumes of digital content. Take, for example, an image archive with a million digital objects. If it takes on average five minutes to compare the two representations, the total process would take approximately 420,000 hours, 53,000 days or 145 days! Clearly we need an automated process to make the evaluation of digital preservation actions practical.

Planets offers a range of services to characterise digital objects:

- Identification services to identify the particular file format of a digital file;
- Validation services to validate whether the content of the digital file conforms to the identified file format specification;
- Extraction services to extract specific characteristics out of a digital file which in turn are used to guide or assess preservation activities;
- Comparison services which build on extraction services to automatically compare specific characteristics of different digital files with each other.

Planets comparison services contain a particularly novel and innovative approach to supporting the automatic comparison of different digital files: Planets **eXtensible Characterisation Language (XCL)**. XCL consists of two parts, XCEL and XDL. **XCEL (eXtensible Characterisation Extraction Language)** is a specification language to define the characteristics of a particular file format and how they should be extracted. . The extraction services will use the XCEL specification to produce the specific results for a particular digital object. These are described using **the eXtensible Characterisation Definition Language (XCDL)**. The XDL produced for two or more digital files is then

automatically compared to decide whether authenticity is maintained between the different versions. Both XCEL and XDL are XML-based.

Currently the characterisation of XCL is focused on the digital content file. Ongoing research will also look at how to include characteristics of the required software stack (processing) and the characteristics of the hardware (presentation).

For further information, please see reading list items C4, C5, D2.

4. How to Preserve

Watch this presentation at:
http://www.planets-project.eu/training-materials/3-van-bussel-how_to_preserve/

Long-term preservation needs to address two aspects - technical preservation, focused on maintaining access to the actual bit stream over time; and logical preservation, making sure that we still can use the bit-stream in a meaningful way. The big challenges are mainly related to logical preservation. Technology is changing very quickly. Every day, new formats and associated programs that can work with the formats are being released. Even for something as common as video or sound, there are a multitude of different formats and associated programs available.

There are two primary approaches to preservation of a digital object for the long-term. Either we migrate the digital object with an obsolete file format to a new current format or we emulate the old software environment (operating system and applications) that support the obsolete file format.

The advantage of **migration** is the ability to use the migrated digital object with currently available software. Normally this adds some additional support like copy and paste functionality inside the complete system. This approach also means that the user doesn't need to know how to operate an old application with an unfamiliar and obsolete interface. However, there are also risks associated with the migration process. Information can become corrupted in the process or essential functionality supported by the original representation application might not be supported in the new environment and the quality of the migration process is difficult to assess. Experiments have shown that consecutive migrations of a digital object can lead to the corruption of the information beyond recognition. For example, a WP5.1 text document becomes completely unreadable through consecutive migration steps to Word 95, Word 97 and finally to Word XP.

Migration actions can be conducted during ingest, at access or inside the electronic repository, depending on institutional objectives. During ingest, certain file formats could already be migrated to formats which are better for archiving, and/or more widely used, i.e. normalisation. Upon access, the digital object could be migrated into a format better suited to the particular user community or target environment; for example, a web browser instead of a desktop publishing program. While stored in the archive, at-risk objects can be migrated to a safer format type.

Emulation uses a different approach by providing an environment in which the original software (operating system and applications) can be used on new hardware technology. One could compare it to current virtualisation approaches where complete systems are virtualised and run on different platforms. The advantage of this approach is that the original digital object does not have to be changed. All the original software is used to render the digital object. However, the effort to emulate an obsolete hardware platform like an old IBM PC or the first Macintosh is not trivial and is a challenging feat in itself. In order for this approach to work, we would also need to establish software archives, which would preserve the old operating systems and applications that will run on hardware emulators. Finally, one would also need to preserve the information on how to use these old applications. There are not many people today that would know how to work with the CPM operating system, or the once-popular WordStar word processor for that platform.

While challenging, it is worth mentioning the increased scalability of emulation, in terms of the number of objects that might be supported as an outcome of a single emulation project. Also, emulation is the most suitable approach, of the two suggested, when it comes to interactive, immersive or time-dependent materials, or for those materials where those three aspects of content, process and hardware are less easily distinguishable (e.g. video games cartridges).

Both migration and emulation have their challenges. Which strategy to follow depends on the particular digital object one wants to preserve and the institution's objectives. An image could simply be migrated to a new format but a popular game can probably only be preserved with emulation. Migration tools themselves are also dependent on a particular hardware/software stack and can become obsolete. Therefore we could also preserve the migration tools themselves through emulation.

Planets has conducted a **gap analysis** to identify the coverage of preservation action tools for the most common digital object types currently maintained in archives, libraries and museums. Among the 76 institutions interviewed, their combined electronic archives contained 107 different file formats. However, half of the total amount of digital objects was made up of only three file formats: TIFF, JPG and PDF. [insert percentage] of the total amount is made up of these formats plus XML, .doc, MP3 and HTML All Planets tools deal with nine of the top ten file formats. This does not mean that the other 104 file formats are not important, but a large group of currently stored digital objects is covered by only three file formats. They may be used extensively in particular communities (eg. cultural heritage, scientific or creative communities). Examples include DAISY, a file format for audio-books, and FITS, a format for sheet music. The analysis found that changes to niche file formats are often supported by their user communities.
The Planets Preservation Action tools are maintained within the **Planets Core Registry (PCR)**. The PCR is based on the Pronom registry developed by the UK National Archives and extends the file format information with information about the required software and hardware to use the file format. The PCR also has links to the Testbed results of particular preservation tools tested within Planets.

Plato, the preservation planning tool developed within Planets, will retrieve both file format and preservation action tool information from the PCR to help define a preservation plan. External users can use the PCR to search for relevant information related to file formats, and finally there is a web service interface to link the PCR to external systems.

For further information, please see reading list item B3.

5. Planets Testbed

Watch this presentation at:
http://www.planets-project.eu/training-materials/5-michaeler-testbed/

Every institution that manages digital collections has to address a number of questions related to the digital collections:

- What file formats do I have in my collection?
- Do I have file formats that are becoming obsolete?
- How do I preserve the obsolete file format objects?

Naturally every institution could address these questions individually, but given the number of file formats and the amount of work needed to address the questions, a better approach is to leverage our joined efforts and experiences. The **Planets Testbed** is created precisely to support this objective, i.e. to leverage and consolidate the knowledge about file formats and associated preservation action tools based on the empirical results of experiments.

Planets has implemented a centrally-managed web-enabled Testbed instance in which experiments can be run and the results centrally stored. The Testbed uses the Planets Interoperability Framework as an open framework on which to base additional preservation services built on web service technology. The individual services could implement a file format identification, a file format validation, extraction of specific properties from the content file or perform an actual preservation action service (migration or emulation).

Currently a number of standard services are already available within the Testbed: JHOVE, PS2PDF, ImageMagik, Sanselan, MSWord migration and HTMLCleaner. The service-oriented integration frame work makes the integration of any other 3rd party tool very easy.

The Testbed also provides classes of a pre-defined set of metrics that can be used to evaluate the performance of a particular service. This way, different services can be compared and assessed in the context of a particular institution. A preservation tool cannot be evaluated without taking the objectives, infrastructure, policy and context of the specific institution into account. A tool that migrates the content of a document but not the layout might be enough for the anticipated user base of institution A but not for B. Therefore the Testbed focuses on the recording of objective individual properties of the services and does not attach any qualitative label to it.

Institutions can use both their own datasets or use existing test sets already stored in the Testbed, i.e. **the Planets Corpora**. Every experiment executes six specific steps:

1. Define the basic properties that will be tested and evaluated during the experiment. For example this could be the speed of the service or the error rate for a specific file format identification service;
2. Design the experiment, i.e. how it will run within the Testbed environment and the activities that need to take place;
3. Specify the resources needed to conduct the experiment;
4. Go/No-go decisions on whether to actually proceed with the experiment based on the experiment set-up and required resources;
5. Run the experiment and gather the results;
6. Evaluate the results.

The 6-step process provides a consistent methodology to evaluate the different preservation services in a controlled hardware and software environment. The available test data makes it possible to run experiments using sample rather than real content and ensure experiments reproducible for everybody. Within the central instance of the Testbed, individual institutions can run their own experiments or look at the results of previously-conducted experiments and make informed decisions. Preservation tool suppliers can test their services in an environment that uniquely identifies the needs of their target customers. Finally, the digital community can use the Testbed as a knowledge base of preservation tools.

One of the great values of a centrally-managed Testbed is the fact that the different experiments are shared across its user community. However, it is also possible for organisations and individuals to set up their own copy of the Testbed and perform experiments locally.

For further information, please see reading list items C8, D3.

6. Preservation Planning with Plato

Watch this presentation at:
http://www.planets-project.eu/training-materials/6-becker-preservation_planning/

Digital preservation is built on trust. Trust encompasses an exchange of values that results in behaviours which conform to expectations. Producers of digital objects must trust the institution that manages the repository to be able to preserve their objects for the long term. Consumers need to trust the deposit organisations to deliver the authentic digital object as the producer intended it. The deposit organisations need to be able to trust the tool providers to supply tools that migrate the digital objects correctly.

The Open Archival Information System (OAIS) Reference Model is often mentioned as a proof of trust for long-term digital repositories. It consists of two high level models outlining information and functional requirements for archiving digital information, but says little about the actual implementation characteristics of a deposit system. Currently a concrete certification process is missing, although multiple initiatives are being introduced. One of the more promising ones is the Trustworthy Repositories Audit and Certification Criteria and Checklist (TRAC)[5]. There are two important aspects of TRAC related to building trust.

1. Repositories need procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve, i.e. policies, plans and monitoring;
2. Repositories also need to document the history of changes to their operations, procedures, software, and hardware where appropriate, link them to relevant preservation strategies and describe potential effects on preserving digital content, i.e. traceability.

The Planets tool Plato helps to build a trustworthy repository by addressing exactly these issues, e.g. preservation plans, monitoring and traceability. At the core is the management of individual preservation plans for specific collections of digital objects. A preservation plan defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records. The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goals. It also describes the preservation context, the evaluated alternative preservation strategies and the resulting decision for one strategy, including the rationale of the decision.

Plato is a web-based tool that provides a four-step process framework to define and manage an institution's preservation plan.
1. The first step is to **define the specific requirements** of a collection of digital objects, e.g. a collection of TIFF images. Plato can help the user to specify the goals and characteristics for that collection of digital objects. Usually these can be defined in at least four major groups: object, record, process and cost characteristics. These are managed in a **requirements** or **utility tree** structure where the leaves are characteristics that are SMART (Specific,

[5] It is worth noting that TRAC, until recently, has not been used for certifying repositories (the Center for Research Libraries (CRL) recently certified Portico in the US on the basis of TRAC). A further development is the MOIMS-RAC work led by David Giaretta to turn TRAC into an ISO Standard via the Consultative Committee for Space Data Systems (CCSDS).

The DRAMBORA toolkit is also relevant when discussing trust and repository evaluation.

Measurable, Achievable, Realistic and Timely). Requirements not defined in a SMART way hamper accurate monitoring of long-term preservation success.

The advantage of Plato is that it already comes supplied with many potential defined requirements that could be used by an institution as a starting point for the requirements specification. Naturally, Plato allows any institution to extend the existing base with new requirements. Many stakeholders may be involved including IT, administration, producers, consumers and curators.

2. Step two is the **evaluation of alternatives**. Different preservation actions are identified using Plato and the Planets Core Registry and the outcomes of actions on an object evaluated against the identified requirements. This is done on a representative test set of digital objects out of the complete collection. The Planets Testbed is used to conduct these experiments and to maintain the results.
3. Step three is the **analysis of the different evaluation results**. Analysis requires criteria to be prioritised, and weighted according to importance, and the results to be transformed into standard values so they are comparable. On the basis of the aggregated results, recommendations are made as to which preservation tool will support the identified requirements for a particular collection of digital objects.

4. On the basis of these recommendations, a specific **preservation plan is defined** for the collection.

A preservation plan is not something that is written once and then not changed over time. Technology innovation and other external influences, for instance budget, will require existing preservation plans to be updated. The knowledge base with all the potential requirements and preservation plans can be shared across different institutions, to leverage existing knowledge. Plato is an important component to set-up a trustworthy repository by explicitly managing requirements and preservation plans, and monitoring them in a traceable way.

For further information, please see reading list items B5, C2, C3, C7, C10, D1.

7. How to Integrate the Components of Digital Preservation with Planets
Watch this presentation at:
http://www.planets-project.eu/training-materials/7-king-how_to_integrate/

The Planets **Interoperability Framework** integrates each of the different components created within Planets. The Interoperability Framework is based on a service-oriented architecture enabling different components to integrate across a web service interface. Digital preservation requires a flexible infrastructure to quickly integrate new services. Every time a new format is introduced, the relevant new identification, validation, characterisation and migration services also have to be introduced into the repository system.

Planets did not set out to build a OAIS-compliant repository system. For instance, it does not have an archival storage component (although there is an assumption that the PCR, Corpora and Testbed results knowledge base will persist). Instead, it provides tools to support the OAIS preservation planning activities within an existing repository system. The Interoperability Framework focuses on providing the environment where different preservation workflows can easily be defined and executed.

A Planets preservation **workflow** is a sequence of Planets services (which are web services that implement one of the specified preservation interfaces such as 'Identify', 'Validate' or 'Migrate'), in which the output parameters of a given service are validly mapped to the input parameters of the subsequent service.

Beside the actual Planets services, there are also generic workflow process definitions for the major processes in the OAIS model: 'Ingest', 'Access' and 'Migration'. They are called **workflow templates**. A workflow template is a workflow in which the nodes (indicating actions to be executed) of the preservation sequence are service placeholders rather than real service implementations. A service placeholder defines only the interface - the actual functionality behind the interface at this stage is irrelevant.

The actual preservation process is defined by a **workflow description**. A workflow description is an XML-serialisation of a Planets workflow, which identifies a workflow template, the service implementations associated with all template placeholders, and the parameters associated with each service. The whole approach is similar to the one defined in a service-oriented architecture with WSDL, BPEL and service bindings in application servers.

The workflow templates provide a quick starting point for an institution to customise their own specific preservation workflow. Take, for instance, the Submission Workflow Template which identifies four basic actions (services) to be executed: validate submission, identify the digital objects (file formats) in the submission package, characterise and validate the digital objects and optionally normalise the valid digital objects (e.g. to PDF/A). Each institution selects the appropriate web service to be used for each of the services described in the workflow template.

The integration to the institution's specific OAIS-compliant digital repository system will take place over the Submission Information Package (SIP) and Dissemination Information Package (DIP) interfaces. One of the challenges of a service-oriented approach is the harmonisation of the data exchanged between individual services. When every service uses a different data model (where there are differences in the interpretation of the supplied arguments and results) the different services will be difficult to integrate because they lack a

shared data model to pass information. The Integration Framework has defined a global **Planets Object Model** to describe the service interfaces. The Data Manager gathers all the relevant data needed to execute the preservation services and convert it into the Planets Object Model for easy exchange of information across the different services.

The Interoperability Framework described above provides the mechanisms needed to link the other Planets components (Testbed, characterisation services, preservation action services, Planets Core Registry and Plato) together to provide a solid and extendible foundation to implement preservation planning in existing digital repository systems.

For further information, please see reading list items C6, C10 plus link to Sourceforge in section D.

Acknowledgements

Planets would like to thank Raymond van Diessen of Consortium partner, IBM, for preparing these technical notes from the face-to-face events.